WS 2017/2018

Graphen und biologische Netze (Modul 10-202-2205; Praktikum)

Project Introduction

Christian Höner zu Siederdissen, Peter Stadler, Thomas Gatter

05.02.2018 - 22.02.2018

1 Introduction

In recent years, RNA-seq technology has become a staple of studies relating to the transcriptomes of both procariotic and eukaryotic organisms [18, 15]. RNA-seq offers a high throughput, low cost methodology for direct sequencing of transcribed genes, thus for the first time enabling the effective identification and quantification of transcript isoforms. Yet, exact classification remains a challenging task, especially for complex eukariotic genes. An increasing number of studies reveals the high degree of diversity in the transcriptomes of higher eukaryotes [2]. Genome wide studies on human tissue revealed that 95% of protein coding multi-exon genes and 30% of non-coding RNAs undergo alternative splicing [12, 3]. Similar ratios have been found also for non-human targets, such as mice [11]. Complicating matters even further, existing annotations likely contain high error rates and are widely considered unreliable.

In a typical RNA-seq experiment, a set of up to 200 mio. paired-end reads is created, each between $100-150$ basepairs (bp) in length. Assembling a set of short reads into a viable set of transcripts is subject to a series of challenges. Transcripts are known to have highly variable sequence coverage, even between isoforms of the same locus. As exons are commonly shared between isoforms, no unambiguous resolutions exist in many cases. In addition to such "natural" concerns, biases from subsequent processing steps or deriving out of the sequencing protocol, have to be considered. As a well known example, both Ribo-Zero as well Poly-A selection will result in deviations in the read distribution [8], as well as other factors such as GC content or position [14, 7]. Therefore, the assumption of reasonably uniform coverage along single isoforms is generally violated. Exon sharing and ambiguous read assignments, e.g. because of close paralogs, and low coverage are additional factors known to hinder quantification [4]. A typical locus with a low level of noise is depicted in Fig. 1.

While an increasing number of methods have been published to solve both the transcript identification as well as the expression quantification problem [13, 17, 16, 9, 1, 10, 5], none of these existing approaches are truly satisfactory in terms of robustness, speed and at the same time accuracy of the results. We devised our own tool Ryūtō that aims for maximal use of evidence from reads spanning more than two exons, as well as of paired-end information. Among its many qualities, Ryūtō is able to show likely areas of errors during the composition of final transcripts.

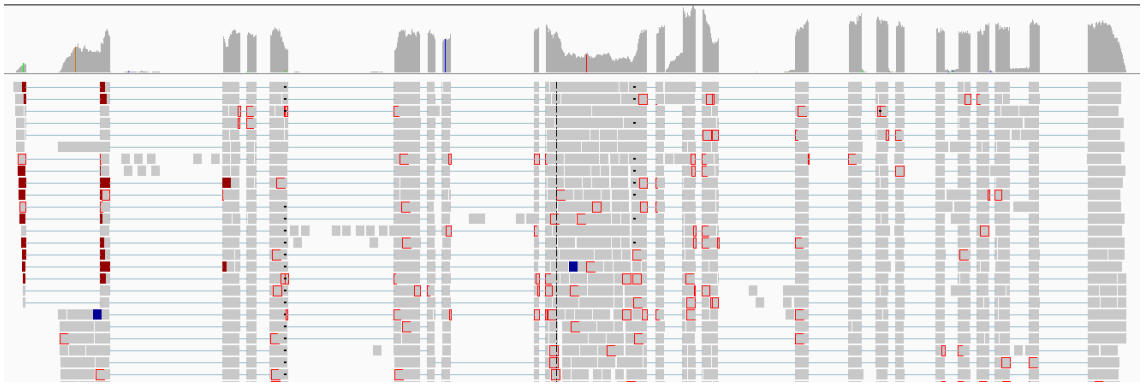

Figure 1: Example for a typical alignment of RNA-seq reads for a single low noise locus. Overlaying isoforms yield signals for different transcripts, e.g. here different start-sites. Noisy reads add false evidence. Coverage varies strongly even within single exons and over the length of possible transcripts. Standard IGV colouring highlights indels and read-pairs with dubious insert length.

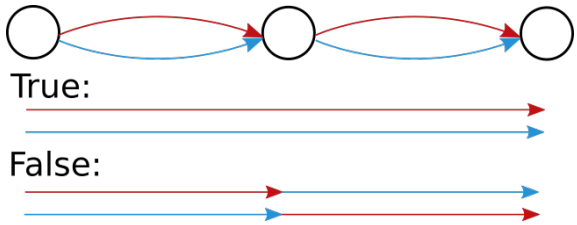

Figure 2: Simplified example for the possible emergence of chimeric transcripts. The connections between left- and right-hand-site of the central nodes (here the colouring) are unknown and have to be estimated by various measures. Transcripts arise as paths through such a graph. When wrong connections are chosen, transcripts can be built entirely from parts of other true transcripts.

These areas arise as the generally short reads can only give evidence for one or only few connected splice-sites. Therefore, the connection between further apart splices often remains unsure and can only be estimated by other (noisy) factors such as crossing coverages. Tools are likely to predict crossed-over transcripts at low evidence areas, e.g. switching starts- and ends between two transcripts or other chimeric transcripts combining parts of actual isoforms in a false manner (see Fig. 2 for an example).

Our method can be simplified to overall 7 steps:

1. First, the individual reads of an RNA-seq experiment are aligned against a known reference genome.
2. Based on these alignments, introns and exons are computed.
3. Each read is assigned to the set of exons corresponding to its alignment. We then compute a generalized splice graph for each cluster of reads. Each splice graph has an artificial source and drain by construction. Transcripts are represented as any unique path from source to drain.
4. After pruning the graph, a flow problem is solved using base-coverages of splice-sites and exons as capacity-limits.
5. Using flow properties, the graph is simplified until only ambiguous nodes (with in- and out-degree > 2) remain in the graph, as well as the artificial source- and drain-nodes.
6. Paired-end, as well as overlapping, reads can be used to further simplify the graph by allowing only paths over adjacent edge-pairs that are directly evidenced.
7. Lastly, decomposing the flow results in a set of transcripts.

Ryūtō will report the ambiguous positions after step 5 is completed by default in its output. However, the usefulness and robustness of this measure remains to be investigated.

During this practical course we aim to explore the properties of these problematic regions. We will use multiple sets of artificially generated RNA-seq reads with known background truth on mouse and human[6]. We want to answer the following main questions:

i Are the ambiguous positions after step 5 and 6 related to properties of the original graph?

```
Exon 12810528 - 12810826 Len 299.
Exon 12810827 - 12810830 Len 4.
Exon 12810831 - 12811580 Len 750.
Exon 12811581 - 12811618 Len 38.
Exon 12811619 - 12825660 Len 14042.
Exon 12825661 - 12825699 Len 39.
Exon 12825700 - 12828628 Len 2929.
Exon 12828629 - 12828714 Len 86.
Exon 12828715 - 12832324 Len 3610.
Exon 12832325 - 12832356 Len 32.
Exon 12832357 - 12832739 Len 383.
Exon 12832740 - 12832761 Len 22.
Exon 12832762 - 12833309 Len 548.
Exon 12833310 - 12833963 Len 654.
Exon 12833964 - 12834026 Len 63.
Exon 12834027 - 12836501 Len 2475.
Exon 12836502 - 12836515 Len 14.
Exon 12836516 - 12836587 Len 72.
Exon 12836588 - 12844519 Len 7932.
Exon 12844660 - 12844661 Len 2.
Exon 12844662 - 12844974 Len 313.
Exon 12844975 - 12846947 Len 1973.
Exon 12846948 - 12848802 Len 1855.
Exon 12848803 - 12849431 Len 629.
Exon 12849432 - 12850261 Len 830.
Exon 12850262 - 12850493 Len 232.
Exon 12850494 - 12850551 Len 58.
Exon 12850552 - 12850609 Len 58.
Exon 12850610 - 12850755 Len 146.
Exon 12850756 - 12850783 Len 28.
Exon 12850784 - 12851069 Len 286.
Exon 12851070 - 12851127 Len 58.
Exon 12851128 - 12851390 Len 263.
Exon 12851391 - 12851500 Len 110.
```

Figure 3: (Sub-)Exon list created by Ryūtō for locus $12,810,528$-$12,851,500$ of mouse chromosome 10.

ii How are ambiguous positions related to the true transcripts? (Did Ryūtō choose wrong connections?)

iii What are the differences between different alignment methods when using the same data?

iv Can we use this information to improve our predictions?

2 Example

An excerpt from section $12,810,528$-$12,851,500$ of mouse (*Mu musculus*) chromosome 10, using simulated data, can serve as an example. As internal representation of splices, Ryūtō uses a bitset of the size of the number of detected exons. In Fig. 3 the (sub-)exon positions of the example region are listed. A bitset of 11000100000000000000000000000000000 in this case indicates a partial transcript over exons $12,810,528$-$12,810,826$, $12,810,827$-$12,810,830$ and $12,825,661$-$12,825,699$. Using this notation on the edges of the splice graph allows for a flexible use at every step. Transcripts are extracted as paths through the graph, combining bitsets with bitwise OR for every edge. Similarly, edge notations are combined by the same operation during simplification steps. In Fig. 4 the basic splice graph is shown. Initially, the graph consists of S-labeled edges, representing splices over potentially multiple exons, and non-labeled edges representing single exons only. This differentiation is a necessary graph-rewrite to allow the use of a standard flow algorithm with capacity limits for both splices and exons. While the initial graph is rather large, it can be reduced to only two critical unresolvable nodes, or even just one node using Paired-end information (Fig. 5). The final output of Ryūtō as well as the base-truth are both represented in GTF format (Fig. 6). In this locus Ryūtō produces both true as well as chimeric transcripts (not shown).

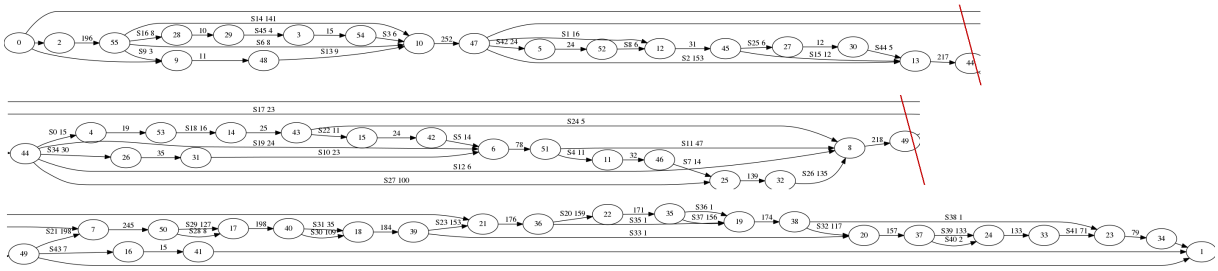

Figure 4: The initial splice graph for locus $12,810,528$-$12,851,500$ of mouse chromosome 10, as created by step 3. Edges are labeled with S-index and coverage. S-Labels: 0 00000000011000000000000000000000000, 1 00000101000000000000000000000000000, 2 00000100010000000000000000000000000, 3 00001100000000000000000000000000000, 4 00000000000000110000000000000000000, 5 00000000000011000000000000000000000, 6 10000100000000000000000000000000000, 7 00000000000000011000000000000000000, 8 00000011000000000000000000000000000, 9 11010000000000000000000000000000000, 10 00000000001001000000000000000000000, 11 00000000000001001000000000000000000, 12 00000000010000001000000000000000000, 13 00010100000000000000000000000000000, 14 11000100000000000000000000000000000, 15 00000001010000000000000000000000000, 16 11100000000000000000000000000000000, 17 00000100000000000001000000000000000, 18 00000000011100000000000000000000000, 19 00000000010000100000000000000000000, 20 00000000000000000000000000110000000, 21 00000000000000000100100000000000000, 22 00000000000011000000000000000000000, 23 00000000000000000000000001100000000, 24 00000000000010000100000000000000000, 25 00000001100000000000000000000000000, 26 00000000000000001100000000000000000, 27 00000000010000001000000000000000000, 28 00000000000000000000111000000000000, 29 00000000000000000000010100000000000, 30 00000000000000000000011100000000000, 31 00000000000000000000001010000000000, 32 00000000000000000000000000111000, 33 00000000000000000000000001000001000, 34 00000000010100000000000000000000000, 35 00000000000000000000000000100100000, 36 00000000000000000000000000010100000, 37 00000000000000000000000000011100000, 38 00000000000000000000000000000100001, 39 00000000000000000000000000000001110, 40 00000000000000000000000000000001010, 41 00000000000000000000000000000000011, 42 00000110000000000000000000000000000, 43 00000000000000000110000000000000000, 44 00000000110000000000000000000000000, 45 00111000000000000000000000000000000

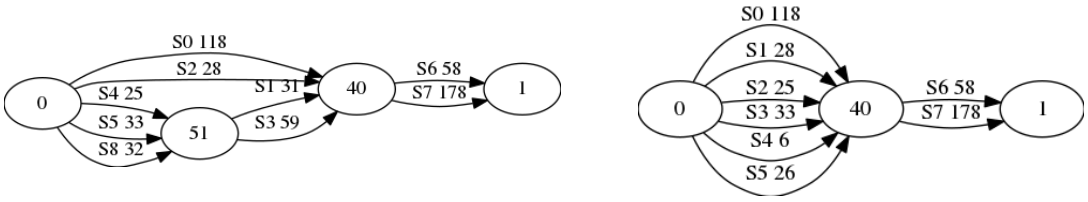

Figure 5: The simplified splice graphs at step 5 (left) and 6 (right). Edges are labeled with S-index and coverage.

S-Labels left: 0 11000100010000001100101000000000000, 1 00000000000001111001010000000000, 2 11000100000000000001010000000000, 3 00000000000000100100101000000000000, 4 11000100011111100000000000000000000, 5 11000100010100100000000000000000000, 6 00000000000000000000000101111111111, 7 00000000000000000000000011111111111, 8 11000100010000100000000000000000

S-Labels right: 0 11000100010000001100101000000000000, 1 11000100000000000000101000000000000, 2 11000100011111111100101000000000000, 3 11000100010100100100101000000000000, 4 11000100010000111100101000000000000, 5 11000100010000100100101000000000000, 6 00000000000000000000000101111111111, 7 00000000000000000000000011111111111

```
chr10 ryuto transcript 12810528 12851500 0 + . gene_id "fres.1_0"; transcript_id "fres.1_0.1"; FPKM "4112.112305";
   read_depth "25";
chr10 ryuto exon 12810528 12810830 0 + . gene_id "fres.1_0"; transcript_id "fres.1_0.1"; FPKM "4112.112305";
   read_depth "25"; unsecurities "14(0), 22(2)";
chr10 ryuto exon 12825661 12825699 0 + . gene_id "fres.1_0"; transcript_id "fres.1_0.1"; FPKM "4112.112305";
   read_depth "25"; unsecurities "14(0), 22(2)";
chr10 ryuto exon 12832325 12836587 0 + . gene_id "fres.1_0"; transcript_id "fres.1_0.1"; FPKM "4112.112305";
   read_depth "25"; unsecurities "14(0), 22(2)";
chr10 ryuto exon 12844662 12844974 0 + . gene_id "fres.1_0"; transcript_id "fres.1_0.1"; FPKM "4112.112305";
   read_depth "25"; unsecurities "14(0), 22(2)";
chr10 ryuto exon 12846948 12851500 0 + . gene_id "fres.1_0"; transcript_id "fres.1_0.1"; FPKM "4112.112305";
   read_depth "25"; unsecurities "14(0), 22(2)";
chr10 ryuto transcript 12810528 12851500 0 + . gene_id "fres.1_0"; transcript_id "fres.1_0.3"; FPKM "5427.988770";
   read_depth "33";
chr10 ryuto exon 12810528 12810830 0 + . gene_id "fres.1_0"; transcript_id "fres.1_0.3"; FPKM "5427.988770";
   read_depth "33"; unsecurities "14(0), 22(2)";
chr10 ryuto exon 12825661 12825699 0 + . gene_id "fres.1_0"; transcript_id "fres.1_0.3"; FPKM "5427.988770";
   read_depth "33"; unsecurities "14(0), 22(2)";
chr10 ryuto exon 12832325 12832356 0 + . gene_id "fres.1_0"; transcript_id "fres.1_0.3"; FPKM "5427.988770";
   read_depth "33"; unsecurities "14(0), 22(2)";
chr10 ryuto exon 12832740 12832761 0 + . gene_id "fres.1_0"; transcript_id "fres.1_0.3"; FPKM "5427.988770";
   read_depth "33"; unsecurities "14(0), 22(2)";
chr10 ryuto exon 12833964 12834026 0 + . gene_id "fres.1_0"; transcript_id "fres.1_0.3"; FPKM "5427.988770";
   read_depth "33"; unsecurities "14(0), 22(2)";
chr10 ryuto exon 12836516 12836587 0 + . gene_id "fres.1_0"; transcript_id "fres.1_0.3"; FPKM "5427.988770";
   read_depth "33"; unsecurities "14(0), 22(2)";
chr10 ryuto exon 12844662 12844974 0 + . gene_id "fres.1_0"; transcript_id "fres.1_0.3"; FPKM "5427.988770";
   read_depth "33"; unsecurities "14(0), 22(2)";
chr10 ryuto exon 12846948 12851500 0 + . gene_id "fres.1_0"; transcript_id "fres.1_0.3"; FPKM "5427.988770";
   read_depth "33"; unsecurities "14(0), 22(2)";
chr10 ryuto transcript 12810528 12851500 0 + . gene_id "fres.1_0"; transcript_id "fres.1_0.5"; FPKM "14474.636719";
   read_depth "88";
chr10 ryuto exon 12810528 12810830 0 + . gene_id "fres.1_0"; transcript_id "fres.1_0.5"; FPKM "14474.636719";
   read_depth "88"; unsecurities "22(2)";
chr10 ryuto exon 12825661 12825699 0 + . gene_id "fres.1_0"; transcript_id "fres.1_0.5"; FPKM "14474.636719";
   read_depth "88"; unsecurities "22(2)";
chr10 ryuto exon 12832325 12832356 0 + . gene_id "fres.1_0"; transcript_id "fres.1_0.5"; FPKM "14474.636719";
   read_depth "88"; unsecurities "22(2)";
chr10 ryuto exon 12836502 12836587 0 + . gene_id "fres.1_0"; transcript_id "fres.1_0.5"; FPKM "14474.636719";
   read_depth "88"; unsecurities "22(2)";
chr10 ryuto exon 12844662 12844974 0 + . gene_id "fres.1_0"; transcript_id "fres.1_0.5"; FPKM "14474.636719";
    read_depth "88"; unsecurities "22(2)";
chr10 ryuto exon 12846948 12851500 0 + . gene_id "fres.1_0"; transcript_id "fres.1_0.5"; FPKM "14474.636719";
   read_depth "88"; unsecurities "22(2)";
...
chr10 feat transcript 12810497 12844666 0 + . gene_id "GENE.56015; transcript_id "GENE.56015"; FPKM "18.4721513926";
chr10 feat exon 12810497 12810826 0 + . gene_id "GENE.56015; transcript_id "GENE.56015"; FPKM "18.4721513926";
chr10 feat exon 12825661 12825699 0 + . gene_id "GENE.56015; transcript_id "GENE.56015"; FPKM "18.4721513926";
chr10 feat exon 12832325 12832356 0 + . gene_id "GENE.56015; transcript_id "GENE.56015"; FPKM "18.4721513926";
chr10 feat exon 12836516 12836587 0 + . gene_id "GENE.56015; transcript_id "GENE.56015"; FPKM "18.4721513926";
chr10 feat exon 12844662 12844666 0 + . gene_id "GENE.56015; transcript_id "GENE.56015"; FPKM "18.4721513926";
chr10 feat transcript 12810518 12844821 0 + . gene_id "GENE.56016; transcript_id "GENE.56016"; FPKM "58.9468520505";
chr10 feat exon 12810518 12810830 0 + . gene_id "GENE.56016; transcript_id "GENE.56016"; FPKM "58.9468520505";
chr10 feat exon 12825661 12825699 0 + . gene_id "GENE.56016; transcript_id "GENE.56016"; FPKM "58.9468520505";
chr10 feat exon 12844662 12844821 0 + . gene_id "GENE.56016; transcript_id "GENE.56016"; FPKM "58.9468520505";
chr10 feat transcript 12810594 12851500 0 + . gene_id "GENE.55310; transcript_id "GENE.55310"; FPKM "81.428898049";
chr10 feat exon 12810594 12810830 0 + . gene_id "GENE.55310; transcript_id "GENE.55310"; FPKM "81.428898049";
chr10 feat exon 12825661 12825699 0 + . gene_id "GENE.55310; transcript_id "GENE.55310"; FPKM "81.428898049";
chr10 feat exon 12832325 12832356 0 + . gene_id "GENE.55310; transcript_id "GENE.55310"; FPKM "81.428898049";
chr10 feat exon 12836502 12836587 0 + . gene_id "GENE.55310; transcript_id "GENE.55310"; FPKM "81.428898049";
chr10 feat exon 12844662 12844974 0 + . gene_id "GENE.55310; transcript_id "GENE.55310"; FPKM "81.428898049";
chr10 feat exon 12846948 12851500 0 + . gene_id "GENE.55310; transcript_id "GENE.55310"; FPKM "81.428898049";
...
```

Figure 6: Example (partial) output of Ryūtō for region for locus $12,810,528$-$12,851,500$ of mouse chromosome 10 (Top) and the underlying truth (Bottom). Both use the GTF format. For Ryūtō, a new tag *unsecurities* is added containing a comma-separated a list of possible sites of errors in the form exon_index(evidence_count). Indexes are relative to the internal exon list also provided as optional output (see Fig. 3).

References

[1] BERNARD, E., JACOB, L., MAIRAL, J., AND VERT, J.-P. Efficient RNA isoform identification and quantification from RNA-Seq data with network flows. *Bioinformatics 30*, 17 (2014), 2447.

[2] BLENCOWE, B. J. Alternative splicing: new insights from global analyses. *Cell 126*, 1 (2006), 37–47.

[3] CABILI, M. N., TRAPNELL, C., GOFF, L., KOZIOL, M., TAZON-VEGA, B., REGEV, A., AND RINN, J. L. Integrative annotation of human large intergenic noncoding rnas reveals global properties and specific subclasses. *Genes & development 25*, 18 (2011), 1915–1927.

[4] GARBER, M., GRABHERR, M. G., GUTTMAN, M., AND TRAPNELL, C. Computational methods for transcriptome annotation and quantification using rna-seq. *Nature methods 8*, 6 (2011), 469–477.

[5] GUTTMAN, M., GARBER, M., LEVIN, J. Z., DONAGHEY, J., ROBINSON, J., ADICONIS, X., FAN, L., KOZIOL, M. J., GNIRKE, A., NUSBAUM, C., RINN, J. L., LANDER, E. S., AND REGEV, A. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol 28*, 5 (May 2010), 503–510.

[6] HAYER, K. E., PIZARRO, A., LAHENS, N. F., HOGENESCH, J. B., AND GRANT, G. R. Benchmark analysis of algorithms for determining and quantifying full-length mrna splice forms from rna-seq data. *Bioinformatics 31*, 24 (2015), 3938–3945.

[7] HUANG, Y., HU, Y., JONES, C. D., MACLEOD, J. N., CHIANG, D. Y., LIU, Y., PRINS, J. F., AND LIU, J. A robust method for transcript quantification with rna-seq data. *Journal of Computational Biology 20*, 3 (2013), 167–187.

[8] LAHENS, N. F., KAVAKLI, I. H., ZHANG, R., HAYER, K., BLACK, M. B., DUECK, H., PIZARRO, A., KIM, J., IRIZARRY, R., THOMAS, R. S., ET AL. Ivt-seq reveals extreme bias in rna sequencing. *Genome biology 15*, 6 (2014), R86.

[9] LIU, J., YU, T., JIANG, T., AND LI, G. Transcomb: genome-guided transcriptome assembly via combing junctions in splicing graphs. *Genome biology 17*, 1 (2016), 213.

[10] LIU, R., AND DICKERSON, J. Strawberry: fast and accurate genome-guided transcript reconstruction and quantification from RNA-seq. *bioRxiv* (2016).

[11] MORTAZAVI, A., WILLIAMS, B. A., MCCUE, K., SCHAEFFER, L., AND WOLD, B. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods 5*, 7 (2008), 621–628.

[12] PAN, Q., SHAI, O., LEE, L. J., FREY, B. J., AND BLENCOWE, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature genetics 40*, 12 (2008), 1413–1415.

[13] PERTEA, M., PERTEA, G. M., ANTONESCU, C. M., CHANG, T.-C., MENDELL, J. T., AND SALZBERG, S. L. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotech 33*, 3 (Mar. 2015), 290–295.

[14] ROBERTS, A., TRAPNELL, C., DONAGHEY, J., RINN, J. L., AND PACHTER, L. Improving rna-seq expression estimates by correcting for fragment bias. *Genome biology 12*, 3 (2011), R22.

[15] SOREK, R., AND COSSART, P. Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. *Nature Reviews Genetics 11*, 1 (2010), 9–16.

[16] TOMESCU, A. I., KUOSMANEN, A., RIZZI, R., AND MÄKINEN, V. A novel min-cost flow method for estimating transcript expression with RNA-Seq. *BMC Bioinformatics 14*, S-5 (2013), S15.

[17] TRAPNELL, C., WILLIAMS, B. A., PERTEA, G., MORTAZAVI, A., KWAN, G., VAN BAREN, M. J., SALZBERG, S. L., WOLD, B. J., AND PACHTER, L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology 28*, 5 (May 2010), 511–515.

[18] WANG, Z., GERSTEIN, M., AND SNYDER, M. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics 10*, 1 (2009), 57–63.